DOCUMENT RESUME

ED 126 151 TM 005 407

AUTHOR Hummel, Thomas J.
TITLE The Analysis of Completely Randomized Factorial Experiments When Observations Are Lost at Random.
REPORT NO ECDO-R&D-R-1976-2
PUB DATE [Apr 76]
NOTE 19p.; Paper presented at the Annual Meeting of the American Educational Research Association (60th, San Francisco, California, April 19-23, 1976)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Data Analysis; *Hypothesis Testing; *Measurement Techniques; Observation; *Probability; *Research Design; Research Problems; Sampling; Statistical Analysis; Statistical Bias; Tests of Significance

ABSTRACT

An investigation was conducted of the characteristics of two estimation procedures and corresponding test statistics used in the analysis of completely randomized factorial experiments when observations are lost at random. For one estimator, contrast coefficients for cell means did not involve the cell frequencies. For the other, contrast coefficients were based on the ratios of cell frequencies to marginal totals. Given the usual normal-theory assumptions and assuming that data lost was unrelated to the treatments, the marginal distributions of estimators and test statistics (across all admissible cell frequencies) were investigated. Marginal means, variances, and probabilities of Type I and Type II errors are presented. (Author)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED126151

THE ANALYSIS OF COMPLETELY RANDOMIZED FACTORIAL EXPERIMENTS WHEN OBSERVATIONS ARE LOST AT RANDOM

by

Thomas J. Hummel

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ECDO R & D Report No. 1976-2

University of Minnesota
College of Education
Education Career Development Office
Minneapolis, Minnesota

April, 1976

(Presented at the 1976 AERA Annual Meeting: San Francisco)

TM005 407

Abstract

This study investigated the characteristics of two estimation procedures and corresponding test statistics used in the analysis of completely randomized factorial experiments when observations are lost at random. For one estimator, contrast coefficients for cell means did not involve the cell frequencies. For the other, contrast coefficients were based on the ratios of cell frequencies to marginal totals. Given the usual normal-theory assumptions and assuming that data loss was unrelated to the treatments, the marginal distributions of estimators and test statistics (across all acceptable cell frequencies) were investigated. Marginal means, variances, and probabilities of Type I and Type II errors are presented.

Acknowledgments

I here express my appreciation to Mr. Charles B. Johnston for writing and running the computer programs used in this investigation and to Professor Raymond O. Collier, Jr. for many helpful discussions and specific suggestions regarding this paper.

Introduction

When an experimenter carries out an experiment it is not uncommon for him to use a frequency in repetitions interpretation of tests of statistical hypotheses and confidence intervals. He might reason that if he uses a particular rule for rejecting the statistical hypothesis, say H: $\Gamma = \Gamma_0$, and if in reality $\Gamma = \Gamma_0$, then the probability of incorrectly rejecting the hypothesis (size of the test) would be ε. (For the moment assume that the size of the test and nominal α are equal.) Probability here would have a "frequency in the long run" interpretation, i.e. if the experiment were carried out over and over again the proportion of incorrect rejections, or "Type I" errors, would be $\varepsilon = \alpha$.

Similarly, when constructing confidence intervals, the experimenter would reason that the random interval would cover the true, or population, value, Γ, in $1 - \alpha$ proportion of the cases. Again, a frequency in repetitions interpretation is commonly used, and the above interpretation of tests and confidence intervals are familiar, of course, to adherents of the Neyman-Pearson approach to testing statistical hypotheses and the confidence interval approach due to Neyman.

Using this type of reasoning, the sample size and configuration (number of observations in the various subclassifications) are assumed to be fixed, e.g. each time a sample is drawn it contains the same number of observations in the same configuration. The set of all repetitions of the experiment, or reference set, which forms the conceptual basis for interpreting probabilities such as the size of the test, therefore, have the characteristic that sample sizes do not change.

Prior to carrying out an experiment the experimenter may not know the

sample size or configuration which will be used when the statistical analysis is performed. He may know that N individuals have been initially sampled, but he may also know that it is likely that some individuals will drop out of the experiment before its completion. Further, certain sample sizes or configurations may be unacceptable to the experimenter and he may be unwilling to analyze the data, perhaps because certain parameters cannot be estimated in a desired manner. If the number of observations remaining in the subclassifications of interest (e.g. cell frequencies) are acceptable to the experimenter, in the sense that the data would be analyzed, one can think of a set of all acceptable sample configurations. Using this concept a new reference set can be defined based on all experiments of a particular type carried out with an acceptable sample configuration.

If there are M acceptable configurations, c_m, $m = 1,\ldots, M$, each occurring with probability p_m, then prior to carrying out the experiment the experimenter will not know which c_m will result, i.e. c_m is a random variable.

If there is some probability of data loss, then the experimenter has two choices. He can define a reference set after the experiment is carried out and c_m is known and therefore can be fixed, or he can define a reference set which allows c_m to be random.

The idea of sample size being a random variable is, of course, not new. In sequential designs a fundamental problem is whether or not to continue the experiment and if a stopping rule is employed, the sample size will be random (see Lindgren, 1968, p. 338 for an introduction).

In 1967, D. J. Bartholomew presented a paper to the Royal Statistical Society which discussed sample size as a random variable, the discussion following being very interesting as well as the paper. In the context of astronomical research, the concept seems implicit in Neyman's (1960) comment that a probability model be defined to account for those comets in the region under study which pass barely out of range or are obscured for some reason.

To further conceptualize this set-up where c_m is random with probability p_m, suppose that the outcome of an experiment, x, has marginal probability density g(x), that $f(x,c_m)$ is the joint probability density, and that $f_m(x|c_m)$ is the conditional density of x given c_m. Attention is commonly focused on $f_m(x|c_m)$, but under some circumstances g(x) may be of interest. From a practical point of view, if experimenter A carries out an experiment he may be concerned about whether or not experimenters B, C, and D will obtain the same results when the experiment is replicated. If $0 < p_m < 1.0$ for all m, then it is not at all certain that all experiments will be carried out with the same c_m and experimenter A may therefore wish to use a reference set in which c_m is random. The random sample size conceptualization also has utility in judging data analysis practices commonly used in educational and psychological research. For example with analyses of variance where data is analyzed with unequal cell frequencies, a probability model for the data loss can be constructed and the performance of various recommended statistical procedures judged. In general, g(x) might be quite difficult (or even impossible) to obtain directly and therefore its relationship to $f_m(x|c_m)$ is important if the characteristics of g(x) are to be

investigated, for in general the $f_m(x|c_m)$ will be known. Now,

$$[1] \quad f_m(x|c_m)\,dx = \frac{f(x,c_m)\,dx}{p_m}, \text{ and}$$

$$[2] \quad p_m f_m(x|c_m) = f(x,c_m)$$

Summing both sides of [2],

$$[3] \quad \sum_{m=1}^{M} p_m f_m(x|c_m) = \sum_{m=1}^{M} f(x,c_m) = g(x)$$

Applying the basic integration rule for the integral of a sum to [2],

$$[4] \quad \sum_{m=1}^{M} p_m \int_{-\infty}^{\infty} f_m(x|c_m)\,dx = \int_{-\infty}^{\infty} g(x)\,dx = 1$$

(The range of summation will be $m = 1,\ldots,M$ and the range of integration will be from $-\infty$ to ∞ unless otherwise denoted).

To this point x has been defined only as an outcome and it may represent a test statistic or an estimate of some parameter. In the former case a critical value k might be of interest, i.e. that value of x which cuts off a tail area equal to α, and in the latter case perhaps $E(x) = \int x\,g(x)\,dx$ and $E(x - E(x))^2 = \int (x - E(x))^2\,g(x)\,dx$ would be of interest. In either case these quantities can be determined by using the weighted sum of the conditional distributions. Further, when testing statistical hypotheses there are a variety of rules which might be applied. One might determine k such that $\int_{-\infty}^{k} g(x)\,dx = \alpha$, as stated above, and use this critical value for all experiments, regardless of c_m. Since the data will always have occurred for a particular c_m, a different critical value may be chosen for each c_m. Perhaps the critical values

would be chosen so that $\alpha = \alpha_m = \int_{-\infty}^{k_m} f_m(x|c_m)\,dx$ thereby ensuring that the marginal α will be controlled at the nominal value even if $g(x)$ is not explicitly considered. If the probability of accepting the hypothesis when it is false, i.e. the probability of a Type II error, or β, is considered, then the critical values k_m might be determined so that the ratio $v\alpha/w\beta$ is constant for all m (here v and w would be weights reflecting the relative importance of the two kinds of errors). The paper by Bartholomew (1967) mentioned earlier gives some discussion of these topics and references which can be used to pursue them.

To this point, the introduction has been quite general without reference to any particular type of experiment or analysis. To apply this general set up, a 2 x 2 completely randomized factorial experiment will now be considered, where the reference set used in determining certain probabilities will be based on all experiments carried out under M acceptable sample sizes and configurations.

The 2 x 2 Completely Randomized Factorial Experiment

In completely randomized factorial experiments, commonly used in educational and psychological research when two or more treatments are employed, each experimental unit (e.g. subject) is randomly assigned to one of t treatment combinations. It is usual to envisage a treatment combination as a joint application of a set of factor levels, one from each treatment factor, so that each treatment combination is applied to, say, n subjects. An example of such a design is a 2 x 2 randomized factorial in which two treatment factors, say, A: caffeine versus no caffeine and B: four hours sleep versus eight hours sleep, are administered, and their effect on reading comprehension is assessed. This type of design is described in many popular texts on experimental design and analysis [e.g. Winer (1962) and Kirk (1968)].

If each of the N = tn subjects (or an equal number of subjects in each treatment combination) completes the experiment, the analysis of the data under the usual normal-theory assumptions is straight-forward and is presented in texts such as those mentioned above. When some subjects do not complete the experiment and the cell frequencies are not all equal, several approaches exist for analyzing the data. The relative merits of various approaches have been discussed recently by Overall and Spiegel (1969), Carlson and Timm (1974), Kutner (1974), and Overall, Spiegel and Cohen (1975).

Suppose that an experimenter chooses a hypothesis to be tested and decides on the analysis to be used before the experiment is carried out. If it is likely that data will be lost at random and that the rate of loss is unrelated to the treatment combination administered, then the

experimenter will have no way of predicting the exact magnitude of cell frequencies. Under these conditions, the experimenter would not want to test a hypothesis which included a function of the cell frequencies, for the hypothesis would contain random variables, i.e. the cell frequencies. While a hypothesis can be specified which would not contain the cell frequencies, distributions of parametric estimates and test statistics conditional on the obtained cell frequencies are, of course, affected by the absolute and relative magnitudes of the cell frequencies such that the experimenter might well be interested in the marginal distribution of estimates and test statistics.

Two estimation procedures and corresponding test statistics used in the analysis of 2 x 2 completely randomized factorial experiments are investigated here. The results, however, are generalizable to more complex designs. Of specific interest is the estimation of the contrast $\Gamma = 1/2\ (\mu_{11} + \mu_{12}) - 1/2\ (\mu_{21} + \mu_{22})$ and tests of the hypothesis $H: \Gamma = 0$. Two estimators are studied: $\hat{\Gamma}_1 \underset{m}{=} 1/2\ (\bar{y}_{11} + \bar{y}_{12}) - 1/2\ (\bar{y}_{21} + \bar{y}_{22})$, to be called the Γ_1 method, or just Γ_1, and $\hat{\Gamma}_2 \underset{m}{=} \frac{1}{n_{1.}}\ (n_{11}\bar{y}_{11} + n_{12}\bar{y}_{12}) - \frac{1}{n_{2.}}\ (n_{21}\bar{y}_{21} + n_{22}\bar{y}_{22})$, to be called the Γ_2 method, or just Γ_2, where $n_{1.} = n_{11} + n_{12}$, $n_{2.} = n_{21} + n_{22}$. Note both n_{ij} and $\bar{y}_{ij}$ are random variables.

Let $y_{ijk} = \mu_{ij} + e_{ijk}$, $i = 1,2$; $j = 1,2$; $k = 1,2,\ldots,n$ where e_{ijk} is $NID(\mu_{ij},\sigma^2)$, be one of the 4n independently distributed observations. Now if $(1 - \pi)$ is the probability that an observation

will be randomly lost, the probability of the m^{th} configuration of cell frequencies is $p_m = Pr[n_{11},n_{12},n_{21},n_{22}|\pi,n] = \prod_{i,j} Pr(n_{ij}|\pi,n)$ where $Pr(n_{ij}|\pi,n) = \binom{n}{n_{ij}} \pi^{n_{ij}} (1 - \pi)^{n - n_{ij}}$, where $m = 1,2,\ldots,M$. (Note that the "m" subscript which n_{ij} and certain other quantities should contain has been dropped for ease of presentation when confusion seems unlikely.) While the probability specification is for cell frequencies ranging from 0 to n, it would always be the case that some configurations would not be of interest. The probability statements excluding unacceptable configurations can be computed by normalizing the probability statement immediately above so that the sum of the probabilities for acceptable configurations equals one. For the m^{th} cell configuration, $\hat{\Gamma}_{1m}$ is $NID(\Gamma_{1m},\sigma^2\ 1/4\ (\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}))$, and $\hat{\Gamma}_{2m}$ is $NID(\Gamma_{2m},\sigma^2\ (\frac{1}{n_{11} + n_{12}} + \frac{1}{n_{21} + n_{22}}))$, where $\Gamma_{1m} = \Gamma$ and $\Gamma_{2m} = \frac{1}{n_{1.}} (n_{11}\mu_{11} + n_{12}\mu_{12}) - \frac{1}{n_{2.}} (n_{21}\mu_{21} + n_{22}\mu_{22})$. The variance ratio statistics for testing the hypothesis H : $\Gamma = 0$ associated with $\hat{\Gamma}_{1m}$ and $\hat{\Gamma}_{2m}$ are conditionally distributed as non-central F-distributions with $df_1 = 1$ and $df_2 = n_{11} + n_{12} + n_{21} + n_{22} - 4$ and non-centrality parameter $\lambda \geq 0$. The marginal density of the estimates and test statistics are weighted sums of the conditional densities of the form $\sum_{m=1}^{M} p_m f_m(x|c_m)$.

It is clear that the expectation of $\hat{\Gamma}_{1m}$, for all m, is $E(\hat{\Gamma}_{1m}) = \Gamma$. In general, however, $E(\hat{\Gamma}_{2m}) \neq \Gamma$.

$E(\hat{\Gamma}_{2.})$, the marginal expectation of Γ_{2m} over all cell frequency configurations, can be written as

$$[5]\quad E(\hat{\Gamma}_{2.}) = \int \hat{\Gamma}_{2.}\, g(\hat{\Gamma}_{2.})\, d\hat{\Gamma}_{2.} = \sum_m p_m \int \hat{\Gamma}_{2.} f_m(\hat{\Gamma}_{2.}|c_m)\, d\hat{\Gamma}_{2.}$$

$$= \sum_m p_m \left[\frac{1}{n_{1.}}(n_{11}\mu_{11} + n_{12}\mu_{12}) - \frac{1}{n_{2.}}(n_{21}\mu_{21} + n_{22}\mu_{22})\right]$$

$$= \sum_m p_m \frac{n_{11}}{n_{1.}}\mu_{11} + \sum_m p_m \frac{n_{12}}{n_{1.}}\mu_{12} - \sum_m p_m \frac{n_{21}}{n_{2.}}\mu_{21} - \sum_m p_m \frac{n_{22}}{n_{2.}}\mu_{22}.$$ But

$\sum_m p_m \frac{n_{ij}}{n_{i.}} = E(\hat{p}_{j|i}) = \frac{p_{ij}}{p_{i.}} = \frac{\pi}{2\pi} = \frac{1}{2}$, and therefore

$$[6]\quad E(\hat{\Gamma}_{2.}) = \tfrac{1}{2}\mu_{11} + \tfrac{1}{2}\mu_{12} - \tfrac{1}{2}\mu_{21} - \tfrac{1}{2}\mu_{22},$$ i.e. the marginal mean is unbiased.

The marginal variance for the Γ_1 method is

$$[7]\quad \sigma^2_{\hat{\Gamma}_{1.}} = \int (\hat{\Gamma}_{1.} - E(\hat{\Gamma}_{1.}))^2\, g(\hat{\Gamma}_{1.})\, d\hat{\Gamma}_{1.}$$

$$= \sum_m p_m \int (\hat{\Gamma}_{1.} - \Gamma)^2 f_m(\hat{\Gamma}_{1.}|c_m)\, d\hat{\Gamma}_{1.} = \sum_m p_m\, \sigma^2_{\hat{\Gamma}_{1m}} = E(\sigma^2_{\hat{\Gamma}_{1m}}) = \sum_m p_m E(\hat{\sigma}^2_{\hat{\Gamma}_{1m}}).$$

Therefore, the variance estimate of $\hat{\Gamma}_{1m}$ unbiasedly estimates $\sigma^2_{\hat{\Gamma}_{1.}}$ since $\hat{\sigma}^2_{\hat{\Gamma}_{1m}}$ is an unbiased estimate of $\sigma^2_{\hat{\Gamma}_{1m}}$ given m, and $E(\sigma^2_{\hat{\Gamma}_{1m}}) = \sigma^2_{\hat{\Gamma}_{1.}}$.

For the Γ_2 method,

$$[8]\quad \sigma^2_{\hat{\Gamma}_{2.}} = \int (\hat{\Gamma}_{2.} - E(\hat{\Gamma}_{2.}))^2\, g(\hat{\Gamma}_{2.})\, d\hat{\Gamma}_{2.}$$

$$= \sum_m p_m \int (\hat{\Gamma}_{2.} - \Gamma_{2m} + \Gamma_{2m} - \Gamma)^2 f_m(\hat{\Gamma}_{2.}|c_m)\, d\hat{\Gamma}_{2.}$$

$$= \sum_m p_m [\sigma^2_{\hat{\Gamma}_{2m}} + (\Gamma_{2m} - \Gamma)^2 + 2\int (\hat{\Gamma}_{2.}\Gamma_{2m} - \hat{\Gamma}_{2.}\Gamma - \Gamma^2_{2m} + \Gamma_{2m}\Gamma)$$

$$f_m(\hat{\Gamma}_2|c_m)\, d\hat{\Gamma}_2] = \sum_m p_m \sigma^2_{\hat{\Gamma}_{2m}} + \sum_m p_m(\Gamma_{2m} - \Gamma)^2 = E(\sigma^2_{\hat{\Gamma}_{2m}}) +$$

$$\sum_m p_m(\Gamma_{2m} - \Gamma)^2.$$

Therefore $E(\sigma^2_{\hat{\Gamma}_{2m}}) \neq \sigma^2_{\hat{\Gamma}_2}$, the variance estimate, $\hat{\sigma}^2_{\hat{\Gamma}_{2m}}$, is a biased estimator of $\sigma^2_{\hat{\Gamma}_2}$ in general.

Both $\sigma^2_{\hat{\Gamma}_{1m}}$ and $\sigma^2_{\hat{\Gamma}_{2m}}$ contain σ^2 as a factor and differ as a function of the cell frequencies. However, $\sigma^2_{\hat{\Gamma}_{1m}} \geq \sigma^2_{\hat{\Gamma}_{2m}}$ for all m, because $\sigma^2_{\hat{\Gamma}_{1m}} - \sigma^2_{\hat{\Gamma}_{2m}} =$

$$[9]\quad \sigma^2[\tfrac{1}{4}\,(\tfrac{1}{n_{11}} + \tfrac{1}{n_{12}} + \tfrac{1}{n_{21}} + \tfrac{1}{n_{22}}) - (\frac{1}{n_{11} + n_{12}} + \frac{1}{n_{21} + n_{22}})]$$

$$= \sigma^2[\frac{n_{11} + n_{12}}{4n_{11}n_{12}} - \frac{1}{n_{11} + n_{12}} + \frac{n_{21} + n_{22}}{4n_{21}n_{22}} - \frac{1}{n_{21} + n_{22}}]$$

$$= \sigma^2[\frac{(n_{11} - n_{12})^2}{4n_{11}n_{12}(n_{11} + n_{12})} + \frac{(n_{21} - n_{22})^2}{4n_{21}n_{22}(n_{21} + n_{22})}] \geq 0$$

Which of the two variances is the larger will depend on the biase factor $\sum_m p_m(\Gamma_{2m} - \Gamma)^2$.

Given an experimental outcome and a particular c_m, on the average (across all m) the researcher using Γ_1 will have an unbiased estimate of error, while one using Γ_2 will on the average underestimate error. On the average, both methods estimate Γ unbiasedly. Considering conventional $1 - \alpha$ confidence intervals, the midpoint of the random intervals will on the average be Γ, however for Γ_2 the intervals will tend to be too narrow

as a function of the bias factor in the variance. If the interval is used to test a statistical hypothesis, say by rejecting H: $\Gamma = \Gamma_0$ when the interval does not cover Γ_0, the Γ_2 interval will tend to reject more often than α proportion of the time when H is true as the bias factor becomes large. It is important to note that when the size of the test with Γ_2, i.e. ε_2, is greater than α, the errors will be symmetrical, i.e. in $\frac{1}{2}(\varepsilon_2 - \alpha)$ proportion of the times one would decide $\Gamma > \Gamma_0$ and in $\frac{1}{2}(\varepsilon_2 - \alpha)$, $\Gamma < \Gamma_0$. More errors will be made using Γ_2 as the bias factor increases, but they will not be misleading in the sense that the excess errors would tend to be in a particular direction.

To translate the above more concretely into the case of the 2 x 2 completely randomized factorial experiment, consider that H: $\Gamma = 0$ represents, say, the "A effect" and also that a "B effect" and "AB effect" are hypothesized. Further consider a table in which H true versus H false is crossed with no B or AB effect versus "large" B and AB effect ("large" in the sense of Cohen (1969)). From the preceding discussion, the table could be filled in with the following relationships (ϕ is power or $1 - \beta$):

	No B or AB Effect	Large B and AB Effect
H: $\Gamma = 0$ True	$\alpha = \varepsilon_1 = \varepsilon_2$	$\alpha = \varepsilon_1 < \varepsilon_2$
H: $\Gamma = 0$ False	$\phi_1 < \phi_2$	$\phi_1 < \phi_2$ (conjecture)

To obtain some idea of the magnitude of the differences $\varepsilon_2 - \varepsilon_1$ and $\phi_2 - \phi_1$, standard numerical methods were used to determine appropriate critical values and to evaluate cumulative distribution functions.

For $\Gamma = 0.0$ or .80 and $\frac{1}{2}\mu_{11} - \frac{1}{2}\mu_{12} + \frac{1}{2}\mu_{21} - \frac{1}{2}\mu_{22} = \frac{1}{2}\mu_{11} - \frac{1}{2}\mu_{12} - \frac{1}{2}\mu_{21} + \frac{1}{2}\mu_{22} = 0.0$ or .80, and $\sigma^2 = 1$, $2 \leq n_{ij} \leq 10$, $\pi = .70$, when H: $\Gamma = 0$ is being tested against H_1: $\Gamma \neq 0$ at $\alpha = .05$, the following results were obtained:

	No B or AB Effect	Large B and AB Effect
H: $\Gamma = 0$ True	$\varepsilon_1 = .05$ $\varepsilon_2 = .05$	$\varepsilon_1 = .05$ $\varepsilon_2 = .06$
H: $\Gamma = 0$ False	$\phi_1 = .51$ $\phi_2 = .52$	$\phi_1 = .51$ $\phi_2 = .52$

From a purely decision making point of view, Γ_1 results in approximately 1% less Type I errors than Γ_2 if B and AB are non-null and large and in approximately 1% more Type II errors. The discrepancies between ε_1 and ε_2, i.e. Type I error rates, are as large as one would obtain for $n > 10$ and/or $\pi > .70$, for as n or π increase the values of $n_{ij}/n_{i.}$ will tend to become less variable and Γ_1 and Γ_2 will yield more similar results. The discrepancies will also tend to decrease as the B and/or AB effects decrease.

Discussion

The appeal of Γ_1 is obvious for one need make no assumptions about π or about the magnitude of contrasts on the μ_{ij}'s other than Γ_1 in order to determine size and power. For Γ_2, size and power will both depend on the value of π and other contrasts. The results would not usually direct one to use Γ_2 as a data analysis procedure.

However, the reference set defined across all acceptable c_m has utility when retrospectively viewing the impact of various analytical procedures on, say, decision making of the kind used when hypotheses are accepted or rejected.

Recently, authors have considered the desirability of testing "weighted" hypotheses (Carlson and Timm, 1974; Kutner, 1974). (A "weighted" hypothesis is of the form $H: a_1\mu_1 = a_2\mu_2 = \ldots = a_t\mu_t$, where $a_1, a_2, \ldots, a_t$ are some function of the cell frequencies.) It seems these authors are of the opinion that some researchers have been unaware of the hypotheses which they "really" have been testing. This point of view would lead one to argue that the type of analysis used to analyze the data from an experiment determines, in fact, the hypothesis which is to be tested. Further, it could be argued that if one used a proper analysis then he would be testing an appropriate hypothesis and be making appropriate decisions. Moreover, if an improper analysis were used, i.e. a method which provided a test of a hypothesis not of interest, then one could be viewed as attempting to answer questions which are of little relevance. In short, the decision making behavior of educational researchers could be far less effective than would be the case if the appropriate

analytical procedures were used.

There is a question as to just how far researchers have been led astray by using "improper" methods of analysis. This question is of importance to educational research and served as an initial motivation for this investigation. By taking the point of view that a researcher determines the hypothesis being tested, not the analysis, i.e. the hypothesis being tested is the one which the researcher will accept or reject, it is possible to consider, say, Type I and Type II errors of competing methods of analysis and/or to study characteristics of competing estimation procedures. This, of course, is the traditional method of approaching this problem. The hypothesis and alternatives are fixed and then the size, power, and possible bias are investigated for various test procedures (e.g. the Neyman-Pearson approach). This approach has a certain common sense, for it seems most reasonable that hypothesis selection preceeds analysis selection and therefore the latter cannot determine the former.

The conclusion to be drawn from the present investigation is that in analyzing data from completely randomized factorial experiments, researchers have not been led far astray. Estimates of effects for Γ_1 are unbiased and error rates for Γ_2 test procedures are close to nominal values. If one is concerned with the "long run" probabilities which would result from replicating an experiment a great number of times, the probability of a Type I error differs little from the nominal α. This might not be the case if all experiments were carried out with the exact same but possibly unequal cell frequencies each time, since substantial bias could be introduced into estimates and tests. But, in general, this will not happen when data is lost in a manner unrelated to the treatments.

As for future practice, as stated above, r_1 does have characteristics which would be desirable to many educational researchers and it seems this technique would be preferable to r_2.

References

Bartholomew, D.J. Hypothesis testing when the sample size is treated as a random variable. Journal of Royal Statistical Society, Series B, 1967, 29, 53-82.

Carlson, J.E., & Timm, N.H. Analysis of nonorthogonal fixed-effects designs. Psychological Bulletin, 1974, 81, 563-570.

Cohen, J. Statistical Power Analysis for the Behavioral Sciences. New York: Academic Press, 1969.

Kirk, R.E. Experimental Design: Procedures for the Behavioral Sciences. Belmont, California: Brooks/Cole Publishing Company, 1968.

Kutner, M.H. Hypothesis testing in linear models (Eisenhart model I). The American Statistician, 1974, 28, 98-100.

Neyman, J. Indeterminism in science and new demands on statisticians. Journal of the American Statistical Association, 1960, 55, 625-639.

Overall, J.E., & Spiegel, D.K. Concerning least squares analysis of experimental data. Psychological Bulletin, 1969, 72, 311-322.

Overall, J.E., Spiegel, D.K., & Cohen, J. Equivalence of orthogonal and nonorthogonal analysis of variance. Psychological Bulletin, 1965, 82, 182-186.

Winer, B.J. Statistical Principles in Experimental Design. New York: McGraw-Hill Book Company, 1962.

Omitted reference, citation on page 3-2.

Lindgren, B. W. Statistical Theory. London: Collier-Macmillan Limited, 1968.